FASA: Frequency-Aware Sparse Attention

[2], Zhenrui Yue[3], Huimin Zeng[3], Yong Wang[1], Ismini Lourentzou[5], Zhengzhong Tu[4], Xiangxiang Chu[1], Julian McAuley[2]
[1]AMAP, Aliaba Group [2]UCSD [3]UIUC [4]Texas A&M University

OVERVIEW

The deployment of Large Language Models (LLMs) faces a critical bottleneck when handling lengthy inputs: the prohibitive memory footprint of the Key Value (KV) cache.

We reveal an overlooked insight into RoPE: the discovery of **functional sparsity** at the frequency-chunk (FC) level. Only a small subset of "dominant" FCs consistently exhibits high contextual agreement with the full attention head.

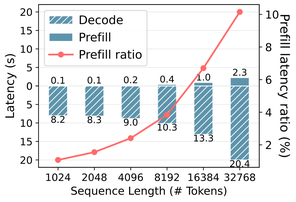

Figure 1: Decoding latency dominates total latency in auto-regressive generation.

This provides a robust and computationally free proxy for identifying salient tokens. Based on this, we propose **FASA**, achieving query-aware token eviction by dynamically predicting token importance without any training.

OBSERVATION & VULNERABILITY

Why functional sparsity exists?

RoPE induces functional sparsity among frequency chunks.

- **Contextual Agreement (CA):** A metric to evaluate how well a single FC's attention distribution matches the full head's attention.

$$\mathrm{CA}_{\mathcal{K}}^{l,h,i}(q_t, \mathbf{K}_{1:t}) = \frac{|\mathrm{TopK}(\boldsymbol{\alpha}_{l,h}, \mathcal{K}) \cap \mathrm{TopK}(\boldsymbol{\alpha}_{l,h}^{(i)}, \mathcal{K})|}{\mathcal{K}}$$

- **Dominant FCs:** A sparse subset of FCs contributes significantly to contextual awareness, while others construct robust positional patterns.

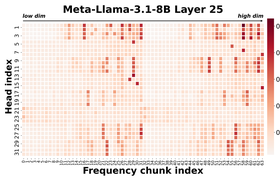

Figure 2: Functional sparsity of FCs revealed by Contextual Agreement ($\overline{\mathrm{CA}}$) heatmaps.

Compound CA Scores ($\mathcal{I}_{dom}$)

$\lvert\mathcal{I}_{dom}\rvert$ \ K	64	256	512	768	1024	2048
Random	2.0	3.6	6.4	19.1	25.5	51.1
Stream	34.4	26.8	24.4	26.5	30.7	53.9
SnapKV	37.9	40.9	41.9	45.4	49.5	66.6
$F = \mathbf{8}$ (1/8)	43.0	49.4	54.3	58.8	62.6	76.1
$F = \mathbf{16}$ (1/4)	**55.3**	**59.7**	**62.8**	**66.9**	**70.1**	**81.4**

Table 1: Compound CA scores under varying selected FCs (F) and budgets (K).

METHODOLOGY: FASA

We employ a two-stage, coarse-to-fine strategy to circumvent the prohibitive cost of full self-attention.

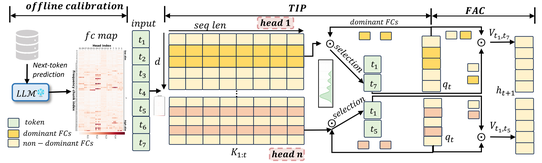

Figure 3: Method Overview of FASA.

1. Token Importance Predictor (TIP)

- **Offline Calibration:** Identify a small, head-specific set of dominant frequencies, $\mathcal{I}_{\text{dom}}$, for each attention head.
- **Online Prediction:** Construct an importance score vector by aggregating contributions *only* from dominant frequencies:

$$\mathbf{S}_t \triangleq \sum_{i \in \mathcal{I}_{\text{dom}}} \boldsymbol{\alpha}^{(i)}(\mathbf{q}_t, \mathbf{K}_{1:t})$$

- Select the top-N_{fac} most important token indices $\mathcal{T}_t$.

2. Focused Attention Computation (FAC)

- Perform full-fidelity attention computation exclusively on the salient subset $\mathcal{T}_t$.
- Original absolute positions are preserved to maintain RoPE embeddings.
- Functions as a high-fidelity computational filter, balancing efficiency and accuracy.

EFFICIENCY ANALYSIS

We introduce two specialized, hardware-aware variants of FASA:

- **FASA-M (Memory-Optimized):** Minimizes GPU memory footprint by offloading value cache and non-dominant key components to CPU. Ideal for VRAM-constrained environments. Provides up to **8× compression**.
- **FASA-C (Computation-Optimized):** Prioritizes inference speed by retaining full cache on-GPU but accessing only a sparse subset of key states. Delivers **2.6× speedup**.

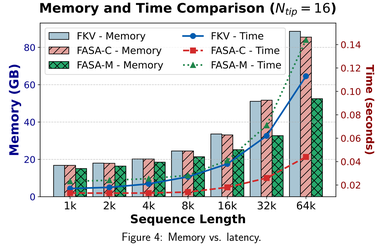

Figure 4: Memory vs. latency.

Complexity & Memory Reduction:

- Speedup $\approx \frac{d}{N_{tip}}$ when $N_{fac} \ll t$.
- Loads only a fraction of the cache: $\approx \frac{N_{tip}}{d}$, heavily alleviating the memory-bound constraint of auto-regressive decoding.

EXPERIMENTAL RESULTS

Long-Context Benchmarks

FASA achieves nearly 100% of full-KV performance on LongBench-V1 when only keeping 256 tokens.

	Method	Single-Doc		Multi-Doc		Summarize		Synthetic		Code	
		NQA	Qasp	Hqa	2Wiki	GovR	Qsum	Pcnt	Pre	Lcc	RB-P
Llama3	FKV	26.0	40.7	32.2	29.6	33.5	22.9	3.5	87.8	52.0	54.2
	Stream	13.2	19.7	18.1	22.7	18.2	17.9	3.5	85.7	49.3	45.9
	SnapKV	23.5	28.9	17.7	22.9	21.7	20.9	3.5	88.0	50.7	48.6
	FASA	**25.6**	**38.9**	**29.7**	**31.2**	**28.0**	**24.2**	3.6	86.9	**53.2**	**50.5**
Qwen2	FKV	24.2	43.5	55.9	46.9	31.8	23.1	7.5	92.0	60.2	66.5
	Stream	18.1	24.2	41.2	36.4	18.3	8.5	24.0	49.6	52.2	
	SnapKV	26.6	36.0	55.6	43.8	21.9	21.9	8.0	98.5	55.6	60.6
	FASA	**28.3**	**43.8**	**57.4**	**46.0**	**31.2**	**22.8**	**8.0**	**99.5**	**60.3**	**64.0**

Table 2: Performance of FASA on LongBench-V1 (Token Budget = 256). FASA matches or exceeds strong baselines.

Language Modeling

Evaluated on PG-19 and Proof-Pile datasets.

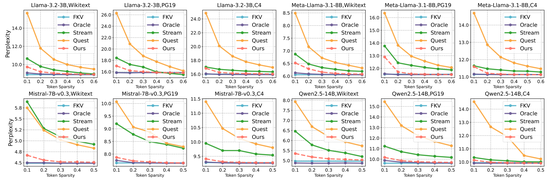

Figure 5: Perplexity comparison on long-sequence modeling. Lower is better.

Long-CoT Reasoning

On AIME24, FASA achieves remarkable robustness even under constrained budgets, showing superior reasoning capabilities compared to standard token eviction methods.

ABLATION STUDIES

Robustness & Trade-offs

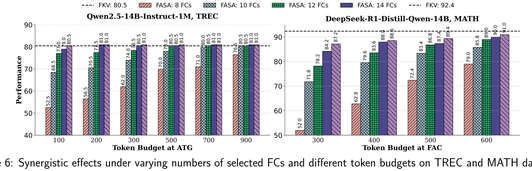

Figure 6: Synergistic effects under varying numbers of selected FCs and different token budgets on TREC and MATH datasets.

Optimal performance can be achieved either with high-precision selection (large N_{tip}) and a small budget, or a more lenient selection compensated by a larger one.

CONCLUSION

- Uncovered functional sparsity at the FC-level induced by RoPE.
- Introduced FASA, a training-free framework for dynamically predicting token importance.
- Consistently achieves near-oracle accuracy across long-context and long-generation tasks while significantly boosting efficiency.